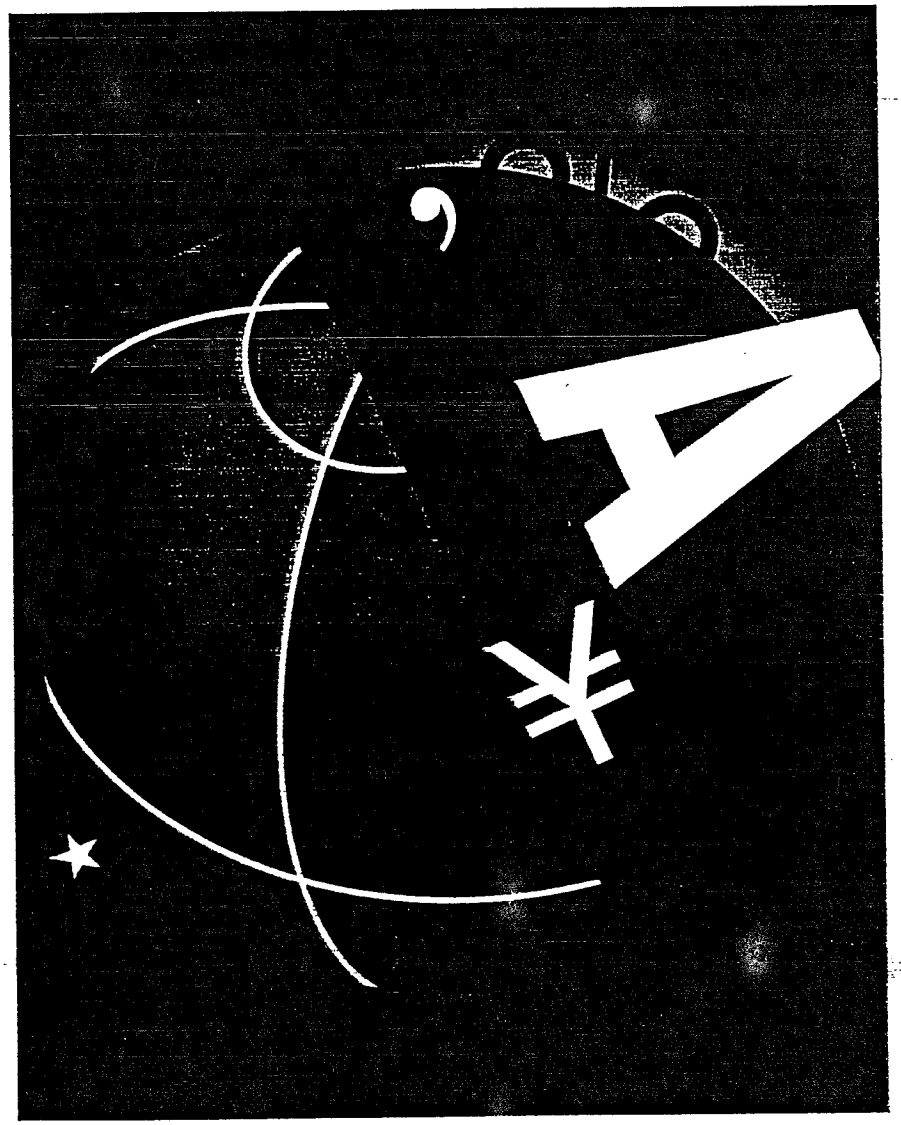
BROWSING THROUGH TERABYTES

Wide-area information servers open a new frontier in personal and corporate information services

RICHARD MARLON STEIN

The Library of Congress archives roughly 25 terabytes in its collection. To browse through this volume on your own would be nearly impossible. Wide-area information servers supply the means to achieve this goal by providing the user-interface structure and underlying information-retrieval protocol necessary to automatically collate, collect, and integrate diverse data streams. WAISes can distill the contents of vast archives into neatly manageable and browsable folders.

On-line information services, such as BIX and CompuServe, attest to the need for this kind of technology. Information has acquired a commodity-like status. While not on a par with wheat, pork bellies, or gold futures, the information-service industry fills a vital role. The next phase of information commerce will add WAIS capabilities to existing on-line services, opening a new frontier in personal and corporate information services.



Intentions and Goals

Initiated in early 1989, the WAIS engineering effort is spearheaded by Thinking Machines (Cambridge, MA), the manufacturer of the Connection Machine, a massively parallel supercomputer (see reference 1). The principal goal of the research project is to demonstrate "how current technology can be used to open a market of information services that will allow a user's workstation to act as librarian and information collection agent from a large number of sources." (See reference 2.) WAISes aim to enhance existing information services and provide a utilitarian mechanism for the industry.

continued

ILLUSTRATION: RON CHAN © 1991

Information servers already provide direct access to many databases and archive structures. You can easily check the local weather, make travel reservations, obtain entertainment schedules, or browse through the latest stock-market quotes on-line. These services are highly interactive, charging users on the basis of minutes spent on-line, and each has a unique user interface.

WAISes alleviate unnecessary user interaction through a predominantly computer-to-computer approach to remote information retrieval. By minimizing human interaction with a remote information server, they handle requests for information expeditiously and inexpensively. WAISes also alleviate unnecessary complexity by moving all user interaction to the local workstation and by having WAIS software handle all transactions with the remote server.

On-line servers are limited in their connectivity. While many services, such as BIX, CompuServe, and AppleLink, incorporate wide-area network structures, sharing information between different services is not a wholly transparent option. This restriction constrains information commerce and hampers the circulation of potentially useful ideas.

WAISes circumvent this barrier with a standard information-exchange protocol that offers unlimited connectivity and retrieval functionality. All servers can apply the WAIS protocol to their archive structures to conduct information retrieval. (Unlimited connectivity also raises concerns of security and privacy. See the text box "The Right to Privacy" on page 160.)

Organized and coherent information of topical importance has value. Individuals and companies should be able to market their information to the widest possible audience. Current on-line services can't easily accomplish this, since their connectivity is restricted.

To direct your information to the best marketplace, you could subscribe to multiple on-line sources and post the same message on all of them. But it would be more efficient to post the data on one server and have the data, or an abstract of it, broadcast to the others. Using the WAIS protocol, WAISes facilitate this server function.

Suppose, for example, you have reviewed the latest set of RISC microprocessor benchmarks, taking note of specific architectural advantages, and you wish to make this information available to others. The benchmark review is kept on your home computer (i.e., the local WAIS), which is equipped with WAIS technology. The nearest remote WAIS, a hub within a network of servers, also has a folder for RISC microprocessors. So you make a posting to the nearest hub server that inserts a pointer to the review on your home computer.

Everyone with a computer running the WAIS user-interface software can present information to a server and receive compensation for whatever portion of it other WAIS subscribers access. The compensation can be monetary, or you can barter your information for someone else's.

Even publishers of books, magazines, newspapers, and music can participate and profit from WAISes. For example, how much money could a newspaper save in circulation costs if you received the morning paper electronically instead of printed on paper? Similarly, how much money could a book publisher save if you purchased a new best-selling novel electronically instead of at a bookstore?

Traditional information delivery is expensive, and costs are rising. The U.S. Postal Service frequently raises its fees to cover increases in the cost of handling and transporting information. Traditional information transport also represents a significant fraction of transport volume and collateral energy consumption. Moving information electronically can result in enormous savings.

Computer networks such as Internet are conduits of information transport. To replace manual transportation methods, the existing electronic infrastructure must accommodate the newly anticipated volume of traffic. Plans for "a national network of data superhighways," which will be installed within the next few years, are under way (see references 3 and 4).

A principal motivation for WAIS technology is to be able to retrieve topical information for research or investigation, not just to deliver consumable items like newspapers or books. Toward this end, WAISes rely on a novel structure for information retrieval, the *dynamic folder*.

To use a WAIS, you formulate a question (see figure 1), find the information servers that provide satisfactory responses, and create a dynamic folder. The purpose of the dynamic folder is to constantly or periodically update its contents with new material on the subject.

Formulating a question is natural to us all. The difficult part is locating the pertinent information to answer it. Manually locating the information can be laborious and tedious. WAISes automate the search-and-retrieval process. To determine which servers hold the information most pertinent to your question, and where you should submit dynamic folders, you may want to consult *server directories*.

Server Directories

WAIS directories are servers that support a directory-services function. They are indexes to other services within the WAIS network and are organized to help you locate information. Like telephone-directory services, WAIS directories list pointers to servers, which are grouped according to content and function.

A *directory-entry header* contains sufficient data to describe the service, such as an English-language description of the server, the parent server (if the server is a subsidiary of a larger one), related servers, contact information (including networks and human-interface points), and cost information.

The local workstation, when equipped with a WAIS, should maintain a directory entry that includes the directory-entry header, a locally determined rank, subscription information (if any), user comments, and the time of last contact. You can use this information to decide whether to contact the server and how to handle the responses.

By using content navigation, you can find the most appropriate server to

BYTE ACTION SUMMARY

The next phase of information commerce will add wide-area information server capabilities to existing on-line services. WAISes provide the user-interface structure and the underlying information-retrieval protocol necessary to automatically collate, collect, and integrate information from various sources. When these are implemented, you should be able to directly access such sources as the Library of Congress and the myriad of newspapers, journals, and books.

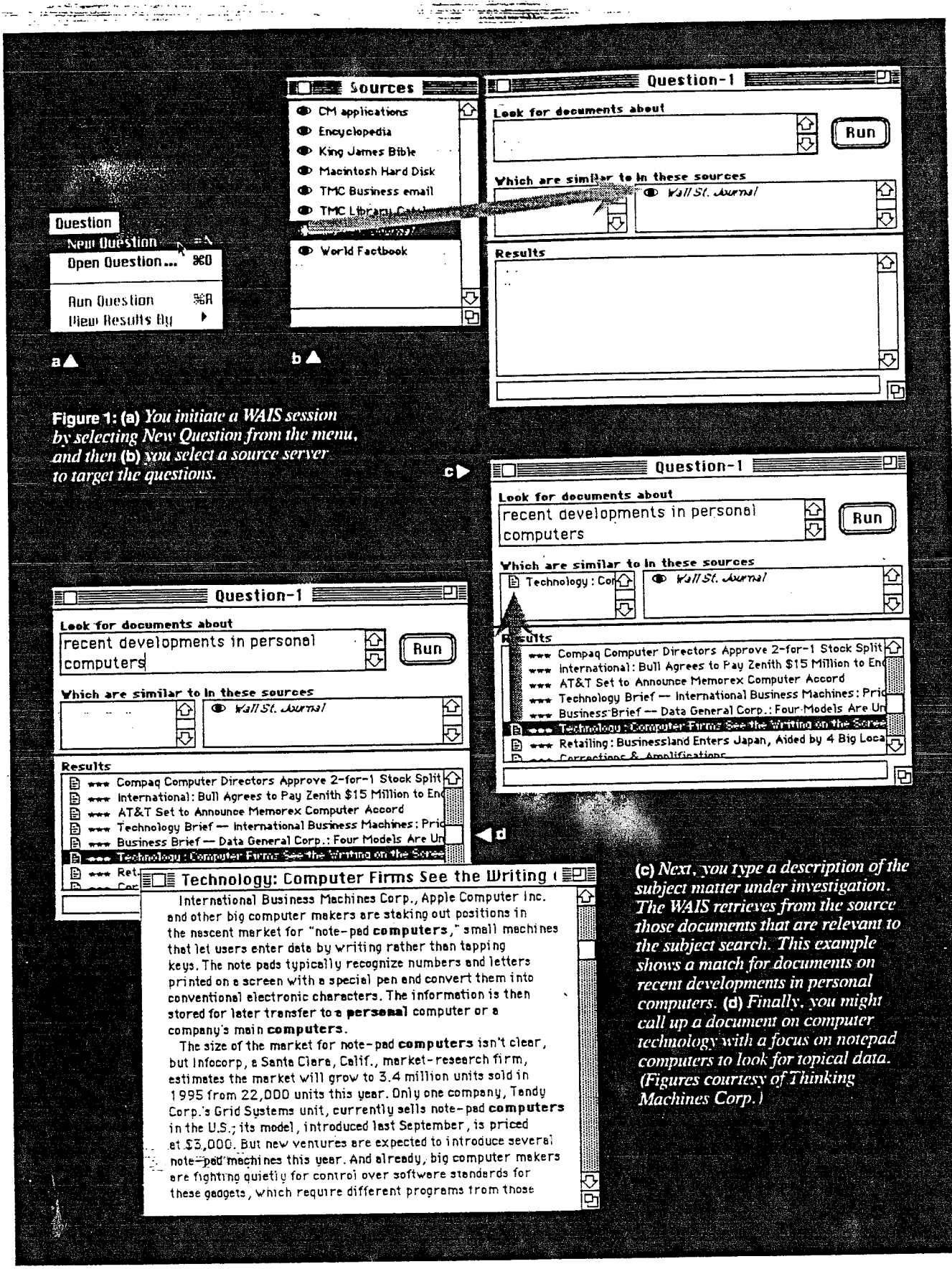


Figure 1: (a) *You initiate a WAIS session by selecting New Question from the menu, and then* **(b)** *you select a source server to target the questions.* **(c)** *Next, you type a description of the subject matter under investigation. The WAIS retrieves from the source those documents that are relevant to the subject search. This example shows a match for documents on recent developments in personal computers.* **(d)** *Finally, you might call up a document on computer technology with a focus on notepad computers to look for topical data. (Figures courtesy of Thinking Machines Corp.)*

The Right to Privacy

WAIStation, a prototype user interface developed by the Thinking Machines wide-area information server project staff, embodies many functional aspects of WAIS technology. Forming and refining queries via relevance feedback, server selection, and dynamic folders are the principal features that this prototype supports. These assets provide a powerful tool set for information retrieval. While WAIStation achieves several desirable technical goals, the security and privacy issues have not yet received serious attention and need refinement.

Security and privacy issues are not specific to WAIStation or WAISes in general, but are endemic, topical concerns of the information-retrieval industry as a whole. WAIS technology seeks to extend connectivity through the WAIS protocol, thus intensifying the urgency of security measures and standards. Greater connectivity promotes information commerce, but it also adds to the risk of compromising the privacy and confidentiality of electronic transactions.

Individuals and corporations that subscribe to WAISes must safeguard proprietary information. The tendency to organize information within a computer for ease of access or to act as a convenient archive creates a security and privacy dilemma. And if the sensitive data is located on a machine with high connectivity, the risk is multiplied.

A WAIStation that holds personal information, such as tax forms, diaries, business transactions, medical records, or bank accounts, must be protected from intrusion by unauthorized individuals. A computer system storing this information "knows" more about you than you can instantly recall. Access to this personal data must be protected, controlled, and limited to authorized individuals.

The WAIS protocol is an application-layer protocol that runs over X.25 communications, modems, or IEEE 802.3 (Ethernet) backbones. Residing beneath this protocol is the WAIStation host computer and operating system. Extracting information from the server depends on access granted through a recognition and authentication system that the host computer operates. Only authorized subscribers can access information from the server.

The WAIS protocol is stateless, so each transaction, whether a query or document-retrieval process, exists in a separate context at the server. Subversion of the WAIS protocol, whether intentional or accidental, might unlock or bypass a server's native file-system protection structure. If it did, the entire archive contents would be available to the intruding party.

The WAIS protocol should be noncorruptible and should detect privileged transactions (i.e., those data streams that possess restricted command sequences). However, to be effective as a noncorruptible application-layer protocol, the underlying computer system must also be unbreachable.

Unfortunately, you cannot always guarantee protection. In 1988, a virus introduced through a known port assaulted computer systems attached to Internet. Subsequent sleuthing discovered that a remote system could activate the debug mode of the Unix mailer, forcing the instigator into a privileged state. The debug mode then permitted the virus to propagate and multiply.

Can a rogue dynamic folder, fashioned after the Internet virus, intentionally access information from strategic servers running WAIS software? How will WAISes safeguard information against illegal intrusion?

The right to privacy is inalienable, and WAIS technology or any enabling system that promotes information commerce must preserve it. A cautionary approach toward implementating WAIS technology is necessary and appropriate. Several legal issues must be addressed to secure both privacy and fair business practice.

handle a query. For example, a question on RISC microprocessor benchmarks would list directory entries for servers as well as pointers to articles on the subject. When you retrieve a document, the directory entry is also provided. Thus, you obtain ranking information for questions of similar content.

Each server, then, contains information of value to certain subscribers. The dynamic folder can continuously poll newspaper servers for new articles as they arrive from the news wires, while it would probably query a dictionary or encyclopedia server only once, since the content changes much less frequently.

Policing the large number of anticipated servers (in the tens of thousands) requires an independent quality-control mechanism. An audit of the server directory would reflect any server that frequently returns erroneous information or does not perform. An independent agency like *Consumer Reports*, the Better Business Bureau, or other watchdog groups could create *rating servers*, which monitor and rate other servers in the directory.

These rating servers resemble movie and TV critics. Consumers acquire confidence in the reports and reviews that certain critics issue because they share similar tastes. Just as moviegoers start to trust a particular reviewer who has agreed with them on past movies, WAIS users will begin to trust the specific rating services that agree with them.

A subscriber base generates income for a server. The rating servers will attract subscribers as well, for they direct trends in the information marketplace. In fact, they may become the first "information speculators" as a by-product of WAIS technology.

Dynamic Folders

A folder, like those found on the Macintosh, provides the WAIS framework for organizing questions. A folder is a repository for documents. A file system, in the Macintosh sense, is full of folders organized in a tree structure that supports an efficient document-location mechanism.

To find a document within a file system, you typically use the find command under Unix or Finder on the Mac.

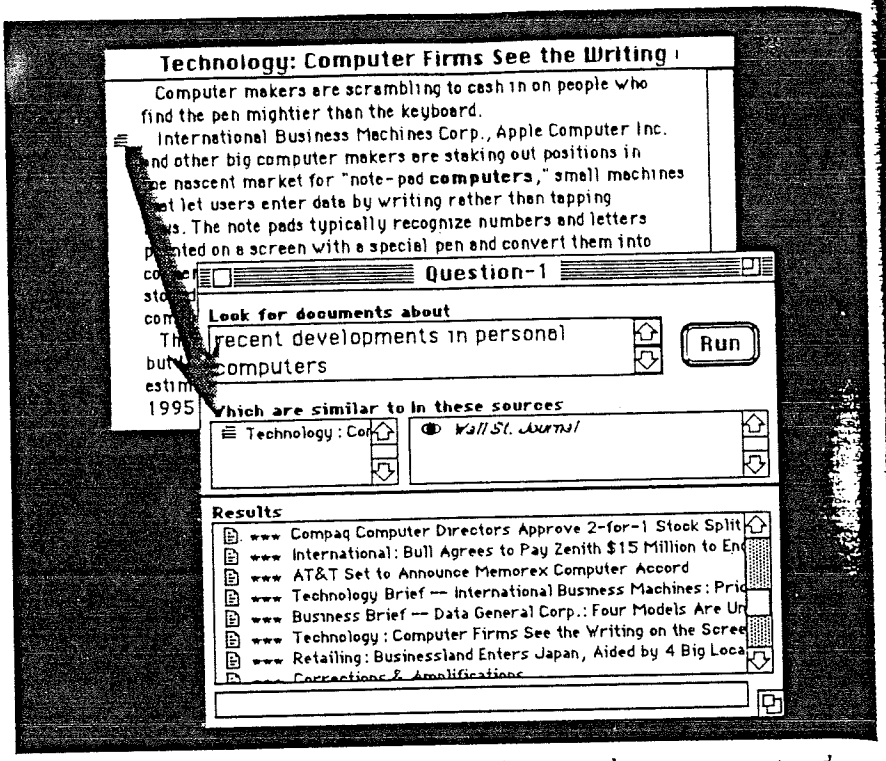


Figure 2: *The* `similar to` *function lets you retrieve more documents on notepad computers using relevance feedback. You then might initiate a search for additional documents with similar content. Selecting text from a section of a retrieved document helps to refine subject-matter searches or locate collateral information. You can also use the selected text to execute a new query. (Courtesy of Thinking Machines Corp.)*

With one of these tools, you can locate the position of a file and gain access to its contents. Path-driven locators search an information base for a document's name, but they do not provide a means to examine its contents.

Retrieving documents pertinent to a specific question requires *content navigation* (i.e., examining the contents of a document, or a representative abstract or index for the document, for its relevance to the question). The similarity between the question and the document's index determines a retrieval score, an indication of the likelihood that the document is pertinent.

WAISes rely on the dynamic folder to encapsulate a question. In its most passive form, it contains a question and a set of servers to target. The WAIS posts the dynamic folder to servers of known quality and functionality, and then query processing begins.

The dynamic folder executes a remote query that sends questions to the remote servers. There the questions find relevant information and return a list of document titles (document pointers) encapsulated within the originating folder to the local WAIS system. The results from the query may initially include a list of documents with fair, good, or high similarities.

Now you can refine your query strategy by perusing the document titles to determine which are the most appropriate documents. WAIS technology, in the form of the WAIStation user interface (see reference 5), assists this process through a content-associativity function known as `similar to`.

The `similar to` function informs the WAIS user interface that a document is "interesting." The server uses this information to find other documents that are similar to the one you have chosen. This search strategy, an embedded component of WAISes, represents a significant improvement over traditional database methods, such as Structured Query Language (SQL) and Boolean search.

This form of query execution is known as *relevance feedback*. It lets you extend the query to incorporate a "more-like-that-one" functionality and lets you retrieve documents that have similar contents. The WAIS user interface is organized around the English language, and English-language-oriented query structures are easier to use than SQL.

The `similar to` function is like working with a reference librarian. First, you state the topic of your research, which the librarian translates into queries. After you examine the results of the queries, you indicate which results were on the mark; thus, the librarian gains a better understanding of your needs and can improve the search.

With relevance feedback, WAISes can retrieve documents with greater ease and speed. You no longer need to alter a SQL Boolean operator to adjust the query filter; instead, you can ask for "more documents like this one."

Dynamic folders can also possess *vitality*, which gives the folder a continuous charter to execute queries periodically and update its contents with new material. A folder's charter expresses purpose, intent, and the goal that you want the query to accomplish. You can build the folder to periodically poll servers known to receive frequently updated material that matches its charter.

If the search retrieves an interesting document, WAISes let you select a portion of the text and use it as an adjunct to the initial query. Selecting text from a portion of a document that may contain some particularly topical or relevant information and using it to refine the search is an innovative approach for exploring subjects (see figure 2).

WAISes also let you chain questions by taking the results of a previous search, starting a new question with different subject matter, and dragging the previous results into the `similar to` menu box (see figure 3). Chaining questions can either broaden or narrow a search, depending on the relevance-feedback results.

The recursive capacity of dynamic folders to initiate "sibling" folders demonstrates the WAIS potential to harness and refine subject matter. Query refinement alters the charter of a dynamic folder. Sibling dynamic folders execute directed searches and can have an autonomous authority to broaden the range of server choices.

Controlling the extent of search expansion is a critical issue. For individuals, cost can be an overwhelming concern. WAIS technology does not yet contain an accounting system to govern search criteria. Participating information services will have to engineer this element of the technology themselves.

WAIS Protocol

WAISes promote connectivity and access to remote electronic-information sources through a standard protocol, the WAIS protocol. This protocol is an extension of the National Information Standards Organization (NISO) Z39.50-1988 specification, which defines an interface to remote information-retrieval services and library-protocol applications. The Z39.50 standard is the backbone of the WAIS protocol and the foundation for WAIS applications development.

With relevance feedback, WAISes can retrieve documents with greater ease and speed.

Incorporating the Z39.50 standard into the WAIS protocol frees developers to build articulated user interfaces for WAIS applications. The interface standard isolates the server's text-retrieval method, such as SQL, giving the application a transparent access mode. The particulars of database queries are hidden beneath the interface. A developer only needs to be sure that the server possesses an equivalent functionality to conduct remote information-retrieval transactions from a local WAIS workstation.

Concealing the server's implementation through the WAIS protocol is important in another respect as well. Isolating the implementation implies that you can specify a single, more palatable query language. The WAIS protocol also lets you use an English-language-style query

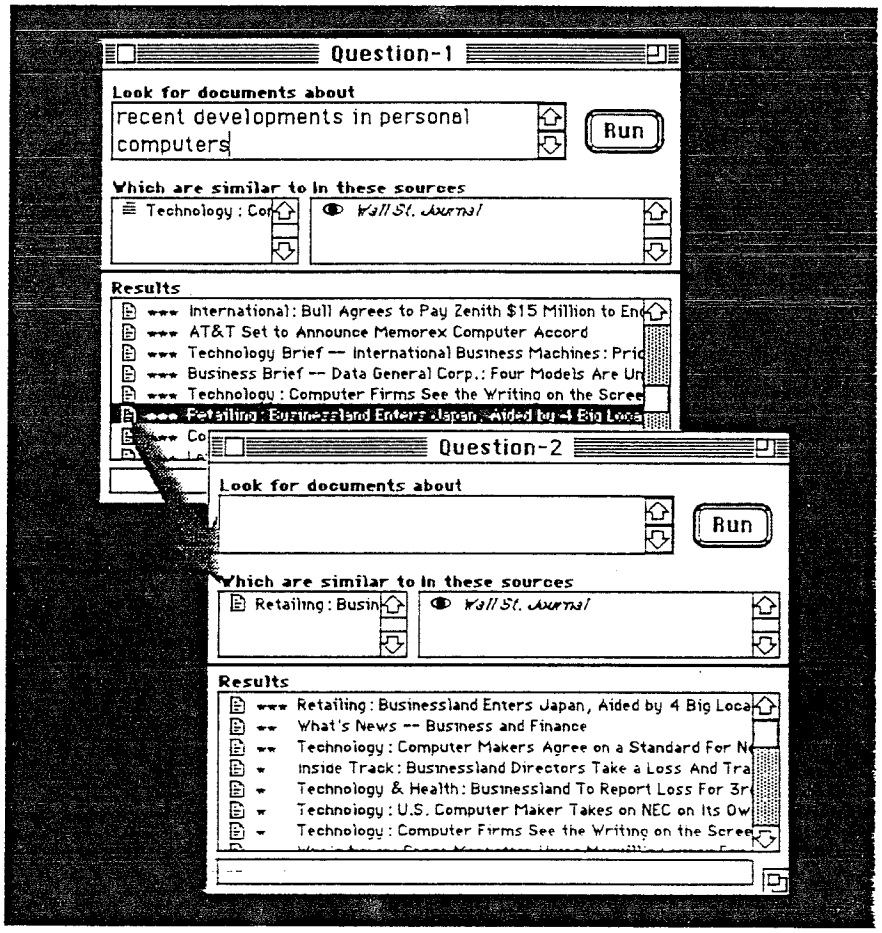


Figure 3: *Chaining questions permits you to use a query on multiple information sources by opening a new question and dragging previous query results into the* `similar to` *field. You can also apply the* `similar to` *operation to invoke a new document search, as in this example. (Courtesy of Thinking Machines Corp.*

lexicon instead of cryptic SQL or fourth-generation languages. When you find a document that is appropriate, the WAIS protocol automatically handles the download process from the server. This is quite different from existing services, where manual file-capture mechanisms require vigilance. With the WAIS protocol, all documents look like they are local to your system.

The WAIS protocol incorporates two important modifications that the NISO Z39.50 standard does not address. First, it permits hypermedia document transport. Most documents today are composed primarily of ASCII text codes and sequences, but the next generation of documents, constructed from hypermedia and multimedia sources, integrates images and fully formatted text. These media forms are rapidly becoming popular and conventional.

While still a research project that is undergoing development and refinement, the WAIS holds immense promise.

Second, the WAIS protocol is stateless for the server. It does not have to keep any information about the client between transactions, because the user's state is kept on the local workstation. Every search or retrieval operation is a separate process. The contexts are decoupled under the statelessness of the protocol. This decoupling lets you make a search, store away the document pointer, and retrieve it later.

Further, you can use a dynamic folder to pass one of these document pointers to someone else who can also retrieve the document. A document pointer is like an International Standard Book Number for the electronic age. (The ISBN is a unique identification assigned to each publication.) Passing a document pointer conforms with copyright law and lets you easily return to the document source instead of making copies.

The WAIS protocol is designed to transport information through modems, X.25 communications, or network backbones. This flexibility provides an enormous framework within which to conduct retrieval transactions. For example, with a portable computer, you could connect with a WAIS hub through a modem and post dynamic folders, directing the query results to be routed to your office system for later examination.

Retrieval Technology

The computing infrastructure needed to implement WAISes varies with a server's functionality. A Library of Congress WAIS, with 25 terabytes of data, could not expeditiously dispatch queries and function if a serial computer were used to process the information. For a problem of this magnitude, massive parallelism is needed. The Connection Machine's Text-Retrieval System is a viable information-retrieval system for gigabyte-size databases.

The DowQuest service from Dow Jones runs on the Connection Machine. The service incorporates approximately 1 gigabyte of original text derived from over 400 sources. The *Wall Street Journal*, the *Washington Post*, *Barron's*, *Fortune*, *Forbes*, and several regional business and technical journals are included, covering the previous eight calendar months. The search time with a 100-word query composed of typed English and relevance feedback (e.g., "more like that one") is less than half a second. The system can provide access to many gigabytes of text and to thousands of users interactively.

The projections for the Connection Machine system indicate that when it is scaled to a 1-terabyte database with 10-word queries, obtaining an answer within 10 seconds or less is highly probable. This performance is accomplished by harnessing the Connection Machine's 65,536 separate processors to execute a parallel index algorithm (see reference 6). These estimates are phenomenal and truly indicative of the computing power manifest in parallel systems. No serial machine can even come close to this level of performance.

The Connection Machine system generates these results by searching the entire contents of an archive, not a representative abstract of a keyword frequency table. Each document within the archive is used to determine a match. This is not typical for systems organized around serial computers, and it is another dramatic demonstration of parallel-computing technology.

The cost of a system like the Connection Machine runs in the millions of dollars. But a Macintosh with a 100-megabyte hard disk drive or a 386-based PC can serve the typical WAIS user.

Immense Promise

The prototype WAIS user interface and protocol are currently being beta-tested at Thinking Machines, Apple Computer, and Dow Jones News/Retrieval. Thinking Machines, the principal developer of the WAIS architecture and software, plans to share the WAIS protocol free of charge and hopes to help user-interface developers build interfaces to WAIS servers.

While still a research project that is undergoing development and refinement, the WAIS holds immense promise. Information commerce, buoyed through the widespread acceptance of computer systems and networks, forces individuals and companies to expedite transactions and simplify activities. These coveted sources of efficiency stand out as prominent allies of competitive advantage. ■

ACKNOWLEDGMENT

I'd like to thank Annie Komanecky, Franklin Davis, Ben Rewis, and Brewster Kahle of Thinking Machines for their assistance during the preparation of this article.

REFERENCES

1. Hillis, D. *The Connection Machine*. Boston, MA: MIT Press, 1985.
2. Kahle, B. "Wide Area Information Server Concepts." *Thinking Machines Technical Memo DR89-1*. Cambridge, MA: Thinking Machines Corp., 1989.
3. Markoff, J. "Computer Project Would Speed Data." *New York Times*, 8 June 1990, sec. A, p. 1.
4. Markoff, J. "Creating a Giant Computer Highway." *New York Times*, 2 Sept. 1990, Part III, p. 11.
5. "WAIStation: A User Interface for Wide Area Information Servers (User Guide, Prototype Version)." Cambridge, MA: Thinking Machines Corp., 1990.
6. Stanfill, C., R. Thau, and D. Waltz. "A Parallel Indexed Algorithm for Information Retrieval." *Thinking Machines Corp. Technical Report DR 90-2*. Cambridge, MA: Thinking Machines Corp., 1990.

Richard Marlon Stein is a software consultant and freelance writer from Van Nuys, California. He has a B.S. in physics from the University of California at Irvine. You can reach him on BIX c/o "editors."